

周报（2013.10.07-2013.10.13）

本周工作：

1. 完成 tour 方法近些年使用情况的调研。详见后面。
2. 完成之前的翻译材料的初稿。

下周工作：

1. 今天和小马交流一下，确定后面需要做的事情。
2. 对翻译进行一次校正。

Tour 方法使用调研：

1. （2）2012 Journal of Statistical Software tourrGui: A gWidgets GUI for the Tour to Explore High-Dimensional Data Using Low-Dimensional Projections
2011 Journal of Statistical Software tourr: An R Package for Exploring Multivariate Data with Projections

这两篇都是 Dianne Cook 做的。这两篇主要是对各种 tour 方法的呈现，包括 grand tour, little tour, local tour, guided tour，投影后的维度可以是 1-6 维等，每种维度可以有不同的呈现方法。

2. （3）2012 TVCG 3D Scatterplot Navigation

研究 3D 散点图的 navigation，在投影插值中保证是刚体运动。在相关工作中提到 tour 方法。提及 grand tour 方法中没有用户交互、在变化中所有维度的投影轴改变了、投影插值过程中不能保证是刚体运动。

3. （2）2011 J Med Syst（医学的，影响因子貌似不是很高，1.783） Data Mining Techniques in Monitoring Diabetes Care. The Simpler the Better?

这篇文章主要用数据挖掘的方法来预测糖尿病患者的某些方面的情况（比如，死亡情况）。主要是比较 6 种数据挖掘方法：Logistic regression (LR), Generalized Additive Model (GAM), Projection pursuit Regression (PPR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Artificial Neural Networks (ANN)。得出的结论是比较简单的模型比如 LR, GAM, LDA 的结果比较好，而 ANN 的结果不是太好。PPR 的结果算是一般吧。

4. （2）2011 International Symposium on Parallel and Distributed Processing with Applications
Network Anomaly Detection based on Projection Pursuit Regression（华中科大的一帮人做的）

将 projection pursuit regression 方法用于网络异常的检测。只是简单地应用。

5. （1.5）2011 Fast Projection Pursuit Based on Quality of Projected Clusters

基于对投影后聚类效果（QPC）的评估的一种 projection pursuit index 通过最小化交叉验证的误差达到了优化投影方向的效果。不过在计算过程中 QPC 需要 $O(n^2)$ 的复杂度，文章提出了一种近似 $O(n)$ 复杂度的方法。引入了一种 a set of prototypes 作为对数据集类别分布估计的参考，介绍这种时用了挺多公式的。

6. （1.5）2011 IEEE TRANSACTIONS ON NEURAL NETWORKS Classifiability-Based Discriminatory Projection Pursuit（中科院的一批人的工作）。

提出了一种新的线性可判别特征提取的计算范式 classifiability-based discriminatory projection pursuit(CDPP)，主要包括两个步骤：构建候选投影集合、寻找可判别投影。

7. （1.5）2011 KDD CHIRP: A New Classifier Based on Composite Hypercubes on Iterated Random Projections

提出了一种新的分类方法，包括三个步骤：projecting、binning、covering。复杂度为次线性的。文中提到 projection pursuit classifiers，指出这种方法的去缺点是计算复杂度大，引用的还是 Dianne Cook 的论文。

8. (1.5) 2011 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing
Trusted Risk Evaluation and Attribute Analysis in Ad-Hoc Networks Security Mechanism based on Projection Pursuit Principal Component Analysis

提出了一种用于风险评估的基于 projection pursuit 的 PCA 方法。测试用的数据集为移动自组织网络。

9. (2) 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics
Risk assessment of real estate investment (河北工程大学的人做的)

将基于实数编码的遗传算法的 projection pursuit 方法用于投资体系的评估。遗传算法用于全局的优化。首先对数据进行归一化，选择一个投影方向，将数据投影到一维空间，采用的 projection pursuit index 为投影后一维数据的标准差和局部密度的乘积。采用遗传算法获得 projection pursuit index 的最优值。

10. (2) 2010 International Conference on Intelligent Computing and Cognitive Informatics
Projection Pursuit Model Based on PSO in the Real Estate Risk Evaluation

用 PSO 算法来优化 projection pursuit index，来构建模型。与第 9 篇论文是相同的人做的提出来的模型也完全一样，唯一的不同点是在求 projection pursuit index 极致时采用的方法不同，上一篇是遗传算法，这一篇是微粒群优化算法(Particle Swarm Optimization, PSO 算法)。两篇文论都不长，3、4 页的样子，同一年的，发在不同的会议上。

11. (2) 2010 Ann Math Artif Intell (影响因子：0.2，稍微有点低)

Genetic algorithms and particle swarm optimization for exploratory projection pursuit (法国人做的，文章 26 页)

这篇文章主要探究了用 Genetic algorithms and particle swarm optimization 求 projection pursuit index 的极值。(呵呵，正好是前面两篇论文里都用到的，不过是完全不相关的人做的)文章提出为了使 Exploratory Projection Pursuit 更强大必须要考虑三方面：采用的 projection index 的适用性、projection pursuit index 的优化、结果的呈现(这个应该就是我们可视化需要考虑的吧)。唯一有点不同的是大部分的优化算法只寻找全局最优值，而这篇文章是用 GA 和 PSO 算法寻找若干个局部最优值(通过多次运行这两种算法获得)，作者关注的也是这些算法处理局部最优值的能力。此外，文章中不仅呈现了一些局部最优值，也把这些局部最优值的频率也进行了显示以及这些投影向量之间的差异。

12. (1.5) 2012 journal of Multivariate Analysis

Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure

文章研究了 kurtosis matrix 的性质，提出了该矩阵的特征向量可以作为一种有趣的方向来揭示数据集的可能的聚类结构。看 introduction 不是很懂，感觉文章的理论性挺强的，里面好多公式。只在 introduction 里面提到了 projection pursuit。

13. (2) 2010 An Efficient Optimization Method for Revealing Local Optima of Projection Pursuit Indices (与第 11 的作者完全一样)这篇文章是基于现有的随机投影方法采用混合 PSO 算法(文中称为 Tribes 方法)优化 projection pursuit index。求解过程中也关注于优化过程中局部最优值的。实验结果表明这种方法用于很大的体数据也比较快。结果也表明一些局部最优值也容易揭示一些有趣的结构。在 conclusion 中也指出 Tribes 方法往往会获得局部最优值而不是全局最优值，在有些人认为这是缺点，不过这反而激发了作者，很好地为他们的目标服务了。

13. (2) 2010 CAR ACO-based Projection Pursuit clustering algorithm (跟 9、10 是同一

个作者，三篇文章还是同一年的)

同样也是对 **projection pursuit index** 的优化，这次采用的是蚁群优化算法 **Ant Colony Optimization** (难道这个作者想把所有的进化算法都用一遍。。。)。这三篇文章提出的模型都完全一样，唯一不同的就是采用不同的优化方法，不过这些方法也完全都是进化算法。这篇文章与另外的两篇相比更理论一些，相当于是提出了一种聚类算法。测试用的数据集也与之前的两篇不同，这次采用的两个测试数据集分别是海水质量状态的统计值和海洋沉淀物的质量。

14. (2) 2010 International Symposium on Intelligence Information Processing and Trusted Computing

A Projection Pursuit Based Risk Assessment Method in mobile Ad hoc Networks (与 8 是同一个作者)

在移动自组织网络的节点的可信性评估时，用 **projection pursuit** 方法，优化 **index** 时采用遗传算法。

15. (2) 2009 International Conference on Electronic Commerce and Business Intelligence

The Research of the credit evaluation Based on Projection Pursuit-Fuzzy Clustering System Model (东北大学的一批人的工作)

将人工智能方法用于信用卡评估研究，采用的方法包括：**projection pursuit** 和模糊聚类 (**fuzzy clustering**)。用 **projection pursuit** 方法来处理训练样本的降维和分类 (用遗传算法获取 **projection pursuit index** 的最优值)，根据最好的投影和分类的结果，用梯形分布方法提取模糊规则，产生三种类型的模糊隶属函数。对于测试样本，根据分布函数和模糊规则，计算模糊近邻以确定样本是否达到信用等级。实验部分使用的数据为中国 105 家上市公司的股票市场数据。

16. (2) 2009 PAKDD

The Effect of Varying Parameters and Focusing on Bus Travel Time Prediction

用回归解决葡萄牙一家公共交通公司的 **travel time of buses** 预测。文章主要是研究参数变化对不同的回归算法性能的影响，也研究了任务聚焦 (样例选择、域值定义、特征选取) 对算法准确性的影响。考虑的算法有 **support vector machine (SVM)**、**random forests (RF)**、**projection pursuit regression (PPR)**。

17. (2) 2009 Seventh International Conference on Advances in Pattern Recognition

The combination of three statistical methods for visual inspection of anomalies in hyperspectral imageries

在高光谱图像的异常检测中应用综合应用三种统计方法 (**PCA**、**the Reed and Xiaoli Yu algorithm**、**projection pursuit algorithm**)。这三种方法在产生异常候选时相互补充，起到了比较好的效果。

18. (2) 2009 ECML PKDD

Subspace Regularization: A New Semi-supervised Learning Method

用半监督学习方法寻找子空间，以便子空间中的决策函数可以很好地分开投影后的数据，并且原始数据的几何信息尽可能地保持。最优子空间和决策函数是通过 **projection pursuit** 迭代寻找的。这种方法的时间复杂度比较低，与其他的半监督方法相比效率也比较高。

19. (2) 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery

Projection Pursuit Model Based on Complex Algorithm and its Application in Water Quality Evaluation of Yuqing Lake Reservoir (山东大学的工作)

提出了一种基于复杂算法 (**complex algorithm**) 的 **projection pursuit model**，用于玉清湖水水质评估。用这种方法可以揭示高维数据的结构特征、水质指标评估结果的不兼容问题。

用复杂算法 (complex algorithm) 寻找最优的投影方向。Projection pursuit index 采用的是投影值的类间散布 (标准差) 和类内强度的乘积。寻找全局最优的 projection pursuit index 采用的是 1965 年的一种非线性规划方法 (complex algorithm)。

20. (2) 2009 IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS

PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach

在交通流缺失数据的补全中采用了一种基于概率 PCA (probabilistic principal component analysis) 方法。用一种 robust PCA 方法过滤出交通流数据中的异常, 以便这些异常不会影响数据补全的过程。Robust PCA 方法是一种结合了 Fast-MCD (minimum covariance determinant) 与 projection pursuit 优点的方法。

21. (2) 2009 Learning Mixed Templates for Object Recognition

提出了一种目标识别的混合学习模板。模板包含局部草图和局部纹理 (分别用于识别目标的形状和外观)。局部草图变量和局部纹理变量的选择采用 projection pursuit 的学习过程。

22. (2) 2009 Second International Symposium on Computational Intelligence and Design

Improved PP model for real estate investment evaluation (与第 9、10、13 是同一个作者的工作, 不过其他的三篇是 10 年的, 这篇是 09 年的)

将 projection pursuit 方法用于投资体系的评估。通篇文章与之前的两篇没什么大的差别, 唯一的区别就是这篇文章中获得 projection pursuit index 最优值的方法是 Ant Colony Optimization (ACO)。个人感觉就是第 13 篇文章就是和独立地提出了这种方法, 而这篇文章是把这种方法用于房地产投资评估。感觉这个作者的这一系列的文章没什么大的创新性。

23. (2) 2009 World Congress on Computer Science and Information Engineering

Genetic Projection Pursuit Interpolation Data Mining Model for Urban Environmental Quality Assessment (北师大的人的工作)

提出了一种 genetic projection pursuit interpolation data mining mode(GPPIDMM)用于城市环境质量的评估。这个模型就是一个最普通的 projection pursuit 方法, 其中 projection pursuit index 最优值通过 genetic algorithm (GA) 方法获得。测试数据集为宣州的环境质量, 主要指标有水环境、大气环境和噪声环境。在 introduction 中稍微提及了 projection pursuit 方法的发展过程: 1974 年最早提出, 用于解决多变量数据的非线性分析问题。

24. (2) 2009 Constructive Neural Networks, SCI

Constructive Neural Network Algorithms That Solve Highly Non-separable Problems 波兰哥白尼大学的工作

(文章挺长的, 22 页, 就看了个大概, 可能还没有完全理解精髓) 提出了一种基于 projection pursuit 方法的构造性神经网络算法, 这种方法可以发现逻辑结构比较复杂的数据中的简单模式。主要的准则是寻找一种变换以发现有趣的结构。Projection pursuit 在 Neural Network 中的应用可以相当广泛, 在网络的隐藏层可以进行各种各样的变换, 或者作为标准神经网络的初始化。文中也提出了一种 Quality of Projected Clusters (QPC, 感觉不是什么高端的东西, 也就是用来评价投影后数据的聚类效果的指标吧) index 用于发现 k-separable 结构以及投影后数据在低维空间中的可视化, 这里的可视化是指投影到二维或者三维空间使得可以对投影后的数据进行可视化, 以揭示隐藏在高维分布中的数据结构。

25. (2) 2009 International Society of Nutrigenetics / Nutrigenomics (ISNN)

An Effective Hybrid GA-PP Strategy for Artificial Neural Network Ensemble and Its Application Stock Market Forecasting (广西柳州师范大学的工作)

提出了一种用于股票市场预测的神经网络集成模型，是一种混合的 genetic algorithm 和 projection pursuit 方法。首先用基于优化的 genetic algorithm 的 projection pursuit 来提取输入因子，也就是把股票数据降维构建神经网络的输入矩阵，然后用 Bagging 方法和不同训练方式获得很多单个的神经网络。然后还是用基于 GA 的 projection pursuit 方法选择适当的集合成员。最后用回归的方法构建神经网络集合。这种方法用于上海股票交易指数的预测，表明其学习能力和扩展能力都比较好。

25. (2) 2009 Complex, Intelligent, and Software Intensive Systems (CISIS)

Agents and Neural Networks for Intrusion Detection

综合计算智能中的三种范式（多重代理系统、基于案例推理、神经网络）用于网络入侵的检测。基于不同统计值的神经模型用于分组式网络交通中的异常检测。还第一次把曲线成分分析的投影方法用于分组式入侵检测。入侵检测中应用的神经投影模型有 PCA、MLHL、CMLHL、CCA。其中 CMLHL 是 MLHL 的一种扩展，而 MLHL 是 a neural implementation of Exploratory Projection Pursuit。通篇只在这里提到了 projection pursuit。

26. (2) 2009 Statistics and Computing

A projection pursuit index for large p small n data (Dianne Cook 的工作)

文章主要探究了分类问题中 the sample size (n) 和 dimensionality (p) 之间的关系，并且提出了一种新的 projection pursuit index (PDA) 用于克服分类问题中的小样本问题。PDA 是 LDA index 的扩展，主要用于 sample size 比较小，data dimension 比较大时投影的选择。在对这个 PDA 优化时采用的是改进的模拟退火算法。PDA 的一个目标是嵌入到 GGobi 中的动态可视化地监测 projection pursuit。作者专注于 p 比较大、 n 比较小的分类性能。测试数据集为基因表达数据和音乐剪辑数据。

27. (2) 2008 International FLINS Conference on Computer Intelligence in Decision and Control Development of Learning Systems with Data Tours Techniques for Fusion Databases

对聚变数据用 grand tour 的方法呈现其分类的结果。最后将这种方法用于 JET（最大的核聚变装置）中的数据检索和中断分类。

27. (2) 2008 Pattern Recognition

A projection pursuit algorithm for anomaly detection in hyperspectral imagery (与 17 是同一个人做的)

提出了一种用于高光谱图像中异常检测的 projection pursuit 方法，基于 Legendre index（因为统计学研究者的经验）的一维 projection pursuit。对该 index 的优化采用模拟退火算法，并且忽略局部最优值。

28. (3) 2007 Robust online signal extraction from multivariate time series

提出了一种鲁棒的基于回归的对多变量时间序列的在线过滤，并且在实时信号抽取设置中讨论其性能。文章只在 reference 中提到了 projection pursuit。

29. (2) 2006 GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization (Dianne Cook 的工作)

提出了一种框架 GGobi，是一种用于交互的数据可视化的从 XGobi 的扩展框架。GGobi 的新特性包括多个绘图窗口、一个颜色查询表管理器、XML 数据格式。GGobi 的最大的进步是可以很容易地扩展（嵌入到其他的软件中或者通过添加插件），此外，也可以通过 API 进行控制。将其嵌入 R 中就是它的一种扩展。

30. (2) 2005 International Society of Nutrigenetics / Nutrigenomics (ISNN)

The SAR Image Compression with Projection Pursuit Neural Networks

将 projection pursuit neural networks 用于 SAR (Synthetic Aperture Radar) 图像压缩。整个过程先把 SAR 图像分割成不同大小的区域，然后用 projection pursuit 的方法为每个

block 构建唯一的编码，当达到期望阈值时停止。

31. (2) 2005 Projection Pursuit for Exploratory Supervised Classification (Dianne Cook 的工作)
由于大部分的 projection pursuit index 在计算过程中不包含分类或者分组信息，并不适用于有监督的分类问题。提出了一种新的 projection pursuit index (源于 linear discriminant analysis) 用于探究有监督的分类，该 index 主要探究 between group variation 相对于 within group variation。
32. (2) 2004 computer statistics
Visualization in classification problems (Dianne Cook 的工作)
描述了可以用于在分类理解数据的聚类结构的可视化方法，即 tour 方法。
33. (2) 2008 Model-driven visual analytics (纽约州立大学的工作)
描述了一个可视化分析的框架，源于机器学习和基于逻辑的演绎推理技术(感觉文章的理论性很强，提到的好多技术我之前没有接触过)。在这里作者描述的是一个用户可以交互地操作的给定的子空间，这个子空间既可以通过 projection pursuit 方法获得，也可以通过 unguided 探究获得。文中提到用 grand tour 时，用户往往会限于观察运动，直到 interesting projection 消失，然而用户交互控制的能力是有限的，不可能以任意方式改变投影方向。在这方面本文的系统提供了更大的灵活性。在相关工作中提到了 GGobi 中的 grand tour 方法。
34. (2) 2007 Tour Generation for Exploration of 3D Virtual Environments
研究 3D 虚拟环境的导航，主要分为两个步骤：1) 离线计算，将世界几何和语意目标信息作为输入，derived from grand tour 2) 在线交互导航，提供一些有指导性的探究和增强的空间感知。前一个步骤是几何数据的一种体素化，用于计算连通图，连通图是用于像 TSP 这种模式的问题，计算环境的 grand tour 以便可以访问到所有重要的 landmarks。而后一步是将前面的输出作为输入。